MathVista: Evaluating Mathematical Reasoning of Foundation Models in Visual Contexts

Year: 2023 | Citations: 1066 | Authors: Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chun-yue Li

Abstract

Large Language Models (LLMs) and Large Multimodal Models (LMMs) exhibit impressive problem-solving skills in many tasks and domains, but their ability in mathematical reasoning in visual contexts has not been systematically studied. To bridge this gap, we present MathVista, a benchmark designed to combine challenges from diverse mathematical and visual tasks. It consists of 6,141 examples, derived from 28 existing multimodal datasets involving mathematics and 3 newly created datasets (i.e., IQTest, FunctionQA, and PaperQA). Completing these tasks requires fine-grained, deep visual understanding and compositional reasoning, which all state-of-the-art foundation models find challenging. With MathVista, we have conducted a comprehensive, quantitative evaluation of 12 prominent foundation models. The best-performing GPT-4V model achieves an overall accuracy of 49.9%, substantially outperforming Bard, the second-best performer, by 15.1%. Our in-depth analysis reveals that the superiority of GPT-4V is mainly attributed to its enhanced visual perception and mathematical reasoning. However, GPT-4V still falls short of human performance by 10.4%, as it often struggles to understand complex figures and perform rigorous reasoning. This significant gap underscores the critical role that MathVista will play in the development of general-purpose AI agents capable of tackling mathematically intensive and visually rich real-world tasks. We further explore the new ability of self-verification, the application of self-consistency, and the interactive chatbot capabilities of GPT-4V, highlighting its promising potential for future research. The project is available at https://mathvista.github.io/.